XVII. Magyar Számítógépes Nyelvészeti Konferencia Szeged, 2021. január 28–29.

Interaktív tematikus-szemantikus térkép a Történeti Magánéleti Korpusz keresőfelületén

Novák Attila[1,2]

[1]Pázmány Péter Katolikus Egyetem, Információs Technológiai és Bionikai Kar
[2]MTA-PPKE Magyar Nyelvtechnológiai Kutatócsoport
Budapest, Práter u. 50/a.
{vezetéknév.keresztnév}@itk.ppke.hu

Kivonat A cikkben a TMK Történeti Magánéleti Korpusz webes lekérdezőfelületének újdonságait mutatjuk be, különös tekintettel a korpusz lexikai anyagát szemléltető, szóbeágyazási modellek felhasználásával készített interaktív tematikus-szemantikus térképekre. A pusztán a TMK-ból készített, a korpusz kis mérete miatt jellegében inkább tematikusnak, mint igazán nyelvinek mondható szóbeágyazási modell mellett a TMK kibővítésével nyert korpuszból készített már inkább nyelvi-szemantikus modellekből a t-SNE algoritmussal nyert kétdimenziós lexikai térképek elemeire kattintva közvetlenül is indítható az adott nyelvi elemre vonatkozó korpuszlekérdezés. A térképek ugyanakkor a szövegek gépi feldolgozásakor, illetve kézi ellenőrzésekor bent maradt hibákra is felhívják a figyelmet, könnyítve ezzel a hibajavítást.

Kulcsszavak: interaktív vizualizáció, t-SNE, szóbeágyazási modellek, korpuszlekérdező, történeti korpusz

1. A Történeti Magánéleti Korpusz

A Történeti Magánéleti Korpusz (TMK)[1] két OTKA, illetve NKFIH kutatási pályázat[2] keretében jött létre a Nyelvtudományi Intézetben. A második pályázat 2020-ban ért véget. A TMK ó- és középmagyar korból származó olyan írott nyelvemlékekből áll, amelyek a magánéleti nyelvi regiszterhez legközelebb állónak tekinthetőek. 1772 előtt keletkezett magánlevelek és perjegyzőkönyvek képezik a korpusz anyagát nagyjából azonos arányban. Elsősorban történeti morfológiai, szociolingvisztikai, történeti mondattani, pragmatikai és lexikológiai vizsgálatokat szem előtt tartva állítottuk össze a korpusz anyagát, és ezek a szempontok határozták meg az annotációs elveket is. A korpusz mérete a második pályázat

[1] `http://tmk.nytud.hu/`

[2] OTKA K 81189: *Morfológiailag elemzett nyelvtörténeti korpusz a magánéleti nyelvhasználat köréből* (2010-2014), NKFI-OTKA K 116217: *Versengő szerkezetek a középmagyar élőnyelvben: változók elemzésén alapuló megközelítés*(2015-2020). Mindkét kutatás vezetője Dömötör Adrienne volt.

zárultakor 8,6 millió karakter, ebből 7,7 millió karakter magyar nyelvű. A magyar nyelvű rész teljes egészében morfológiailag annotált, ez összesen 1 millió 112 ezer elemzett szövegszó.[3]

A korpusz nyomtatott forráskiadások feldolgozásával készült.[4] A szövegek eredeti alakját a nyomtatott kiadásokban szereplő formában vettük át. Ez a szövegváltozat OCR-ezéssel és az így digitalizált szöveg kézi javításával állt elő. A szövegeket félautomatikus módon tagmondatokra bontottuk, majd a tagmondatokra bontást kézzel javítottuk. A más tagmondatokba beágyazott tagmondatokat külön megjelöltük. A tagmondatokra bontott szövegekhez kézzel a mai magyar helyesírásra normalizált változat készült. A normalizálás során neutralizáltuk a morfológiai következményekkel nem járó tisztán fonológiai jellegű nyelvjárási sajátosságokat, de nem változtattunk a szavak morfológiai szerkezetén: a történeti szövegekre jellemző morfológiai és szintaktikai szerkezeteket a normalizálás nem érintette.

A szövegeket a Humor morfológiai elemző (Novák, 2003) ó- és középmagyar szövegekre adaptált változatával (Novák és Wenszky, 2013) elemeztük morfológiailag, és a PurePos szófaji egyértelműsítő eszközzel (Orosz és Novák, 2013) egyértelműsítettük automatikusan. A géppel elemzett és egyértelműsített szövegeket egy erre a célra készült webes egyértelműsítő felületen manuálisan ellenőriztük és javítottuk. Itt az esetleges elemzési illetve egyértelműsítési hibákon kívül a normalizálási, tokenizálási és tagmondatokra bontási hibákat is javítani lehet, és a javításuk után a javított részeket újra lehet elemeztetni. A kézi ellenőrzésen és javításon átesett szövegek a projekt előrehaladása folyamán folyamatosan bekerültek a PurePos egyértelműsítő tanítóanyagába. A Humor morfológiai elemző lexikonját is folyamatosan bővítettük az újonnan elkészült normalizált szövegek szóanyagával. A projekt folyamán nem került sor a korpusz teljes elemzett anyagának kézi ellenőrzésére: az elemzések 78%-a van kézzel ellenőrizve.[5]

A szociolingvisztikai szempontú kutatások segítése érdekében minden szöveget annotáltunk a rendelkezésre álló metaadatokkal. Ezek között minden esetben megtalálható az adott szöveg keletkezésének éve, illetve a levelek esetében pontos dátuma, a keletkezés helye, perek esetében a megye és a műfaj (levél, illetve per). A levelek esetében emellett a szerző, illetve a címzett neve, neme, illetve társadalmi státusza, a szerző és a címzett közötti viszony jellege, valamint az adott szövegrész saját kezű mivoltára vonatkozó információ szerepel a metaadatok között. Ezen kívül az egyes szövegrészeket annotáltuk a szövegrész típusa szerint a szövegtörzs mellett megkülönböztetve a címzést, a külzetet, a margón tett megjegyzéseket és a mellékleteket, illetve perek esetében a formulaszerű hivatalos részeket.

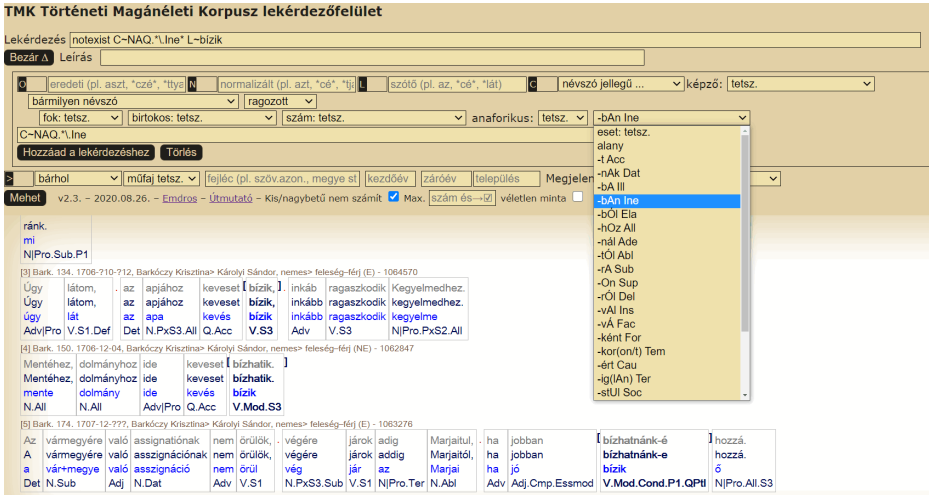
[3] Ez valóban ennyi szót és nem tokent jelent, az írásjeleket nem tekintettük külön tokennek.

[4] `http://tmk.nytud.hu/forrasok.php`

[5] Korábbi méréseink során (Dömötör és mtsai, 2017) a gépi egyértelműsítés pontossága a szótokenek szintjén az írásjeleket nem figyelembe véve 95,9%-osnak, a tagmondatok szintjén 81,5%-osnak adódott (a tagmondatok ötödében találtunk hibát).

2. A TMK webes felülete

A korpusz a `http://tmk.nytud.hu/` címen elérhető. A korpusz keresője az Emdros korpuszkezelőn alapul (Petersen, 2004). Ez lehetővé teszi a korpuszt alkotó szövegek hierarchikus szerkezetének ábrázolását, és a megfelelő részek metaadatokkal való annotálását (pl. a többszerzős levelek megfelelő részei is annotálhatók az adott szövegrész szerzőjével), illetve az ezekre vonatkozó szűrések illetve lekérdezések kezelését. Emellett lehetőséget biztosít a megszakított tagmondatok kezelésére is: alapesetben a beágyazott tagmondatok tartalmát nem tekinti a megszakított tagmondat részének, így az egy tagmondatra korlátozódó lekérdezések mindig helyes eredményt adnak. A korpuszhoz egy az Emdros viszonylag körülményes MQL lekérdezőnyelvénél sokkal tömörebb és egyszerűbben használható korpuszspecifikus keresőnyelvet alakítottunk ki, emellett a lekérdezések összeállításának segítése érdekében lekérdezésszerkesztőt hoztunk létre a kereső webes felületén (1. ábra).[6]



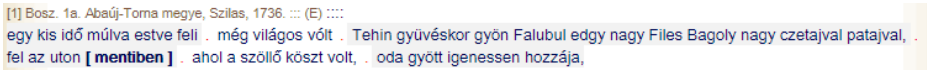
1. ábra. A TMK lekérdezőfelülete a lekérdezésszerkesztővel

A lekérdezések alapesetben egy tagmondaton belüli szavakra tett megszorításokból állnak, amelyek a szó eredeti és normalizált alakjára, annak szótövére és a morfoszintaktikai annotációjára vonatkoznak. A lekérdezésszerkesztő mindezen tulajdonságokra vonatkozó megszorítások megfogalmazásához segítséget nyújt. A morfológiai jellemzők egy hierarchikusan automatikusan kibomló menürendszer segítségével választhatók ki. Az 1. ábrán látható helyzetben a névszói esetrag kiválasztása látható a lekérdezésszerkesztő segítségével, a teljes lekérdezés pedig azt írja le, hogy olyan tagmondatokat keresünk, amelyekben a *bízik* lemmájú ige mellett nem szerepel inesszívusz esetű névszó.

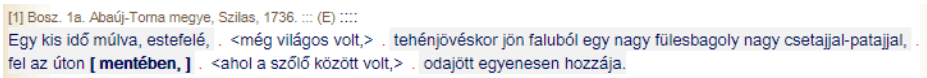
[6] Részletesebben l. `http://tmk.nytud.hu/utmutato.php`.

Alapesetben a találati egységek mondatok, amelyek tagmondatokra vannak bontva és a találatot adó szavak ki vannak emelve. Ez a kiemelés az Emdros terminológiájában a *fókusz:* a példában a *bízik* alakjai. Alapesetben a mondatok interlineáris formátumban jelennek meg (1. és 2a ábra) és a szavak eredeti és normalizált alakját, szótövét és a morfoszintaktikai annotációját külön sorokban tartalmazzák. A megszakított tagmondatokat eltérő háttérszín jelzi. Minden egyes mondattalálat fölött szerepelnek a találatot adó szöveg főbb jellemzői. A szövegazonosító mellett a dátum, szerző és címzett, illetve a per helyszíne, a szerző és a címzett viszonya (az 1. ábrán levelekből, a 2. ábrán perszövegből származó találatokat látunk). Itt jelezzük emellett, hogy az adott szöveg átesett-e a gépi annotációt követő kézi ellenőrzésen (E=ellenőrzött, NE=nem ellenőrzött). A találathoz tartozó metaadatokra kattintva külön ablakban a teljes szöveg megnyílik, amelyen belül a keresésben találatot adó szavak ugyanúgy ki vannak emelve, mint az eredeti egymondatos találatokban. A teljes annotáció mellett a találatok egyszerűsített formában morfológiai annotáció nélkül is megjeleníthetőek. Ebben a változatban választható, hogy a találatokat az eredeti (2b ábra) vagy a normalizált alakjukban szeretnénk látni (2c ábra). A teljes mondatos találatok mellett gyakorisági adatok is kérhetők a rendszertől. Ilyenkor megadható, hogy a találati elemeknek melyik jellemzői jelenjenek meg.

(a) Interlineáris megjelenítés - beágyazott tagmondatokkal



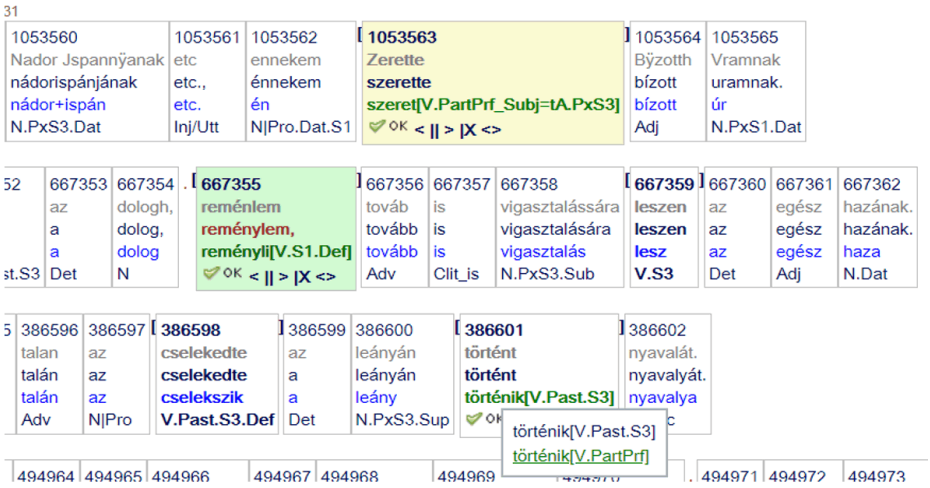
(b) Egyszerűsített megjelenítés - eredeti alak



(c) Egyszerűsített megjelenítés - normalizált alak

2. ábra. Megjelenítési formátumok a korpuszlekérdezőben

A kereső speciális szolgáltatása, hogy a megfelelő jogosultsággal rendelkező felhasználók számára lehetővé teszi a keresőben való hibajavítást is (3. ábra). Egy adott szóra kattintva a kézi egyértelműsítő felülethez hasonló módon javítható a szó eredeti, illetve normalizált alakja, elérhető a morfológiai elemző, melynek elemzései közül választhatunk, illetve kézzel is szerkeszthető az elemzés. Emellett a tokenizálási és tagmondatokra bontási hibák javítására is van lehetőség.



3. ábra. Hibajavítás a TMK kereső találataiban.

3. Szemantikus térképek a korpusz lexémáiról

A projektum zárószakaszában a keresőt egy új szolgáltatással egészítettük ki. Ez neurális disztribúciós modelleken alapuló kétdimenziós lexikális térképekből áll, amelyek a korpuszban legalább háromszor előforduló lexikai elemek disztribúciós szemantikai térben való reprezentációját vizualizálják.[7] A egyes elemek a gyakoriságukkal (logaritmikusan) arányos méretben és a szófajukra jellemző színben jelennek meg (5., 7., 8. ábrák). A térképböngészőbe keresési funkciót is integráltunk, amelynek segítségével a térképen szereplő lexikai egységekre kereshetünk illeszkedő szórészletek alapján. A találatok átmenetileg kiemelt színnel és kinagyítva jelennek meg, illetve egyenként végiglépegethetünk rajtuk az adott elem környékére automatikusan ráközelítve. A térképek annyiban interaktívak, hogy a rajtuk szereplő lexikai elemekre duplán kattintva lekérdezés kezdeményezhető az adott elemre a korpuszból. A lekérdezés eredménye új böngészőfülön jelenik meg (9. ábra).

3.1. Előzmények

Korábban milliárdszavas nagyságrendű magyar nyelvű webkorpuszból hoztunk létre a word2vec (Mikolov és mtsai, 2013), illetve a fastText (Bojanowski és mtsai, 2016) eszköz CBOW modelljével háromszáz dimenziós disztribúciós modelleket. Nyers szövegen tanított modellek mellett morfológiailag annotált szövegen betanított modelleket is létrehoztunk, amelyek a ritkább szavakra jobb minőségű reprezentációt hoztak létre, mert a lemmatizálás csökkentette az adatritkaságot

[7] `http://tmk.nytud.hu/maps.php`

(Novák és Novák, 2018). Azokban a modelljeinkben, amelyekben a fő szófajcímkét is a lemmatizált elemek részévé tettük, a módszer azon hiányosságát is sikerült részben kiküszöbölni, hogy önmagában nem alkalmas a homonímia, illetve poliszémia kezelésére.[8] A modelljeinket korábban t-SNE (t-distributed stochastic neighbor embedding) algoritmus (van der Maaten és Hinton, 2008) segítségével vizualizáltuk és jelen kutatásban is ezt a módszert alkalmaztuk.

A korábban létrehozott sok millió lexikai elemet tartalmazó modelljeink esetében a vizualizációt a modellt böngésző felhasználó által menet közben összeállított korlátos szókészletre dinamikusan hoztuk létre a szerveren (Novák és mtsai, 2017). Mivel a t-SNE algoritmus gradiens ereszkedés algoritmussal (SGD) optimalizálja a képet eloszlások Kullback–Leibler (KL)-távolságát hibafüggvényként használva,[9] ezért futtatása a szerveren meglehetősen idő- és erőforrás-igényes (sok ezer pont megjelenítése esetén a keresést futtató szerveren több percig tarthat az ábra generálása). Ezt a TMK keresőfelületére integrálandó interaktív vizualizáció esetében mindenképp szerettük volna elkerülni. Korábban kísérleteztünk autoenkoderen alapuló vizualizációval is, amely a képgenerálás válaszidejét jelentősen csökkenthetné, ez azonban a szóbeágyazási modellen alapuló szemantikus térképek megjelenítésére nem adott elfogadható minőségű megoldást (Novák és Novák, 2020).

3.2. A TMK felületén alkalmazott megoldás

Ugyanakkor a szemantikus térkép megjelenítése a kliens gép böngészőjében szintén túlzott erőforrásigényt jelent, ha a modell túl nagy. Ez a modellt nézegető felhasználó gépén a böngésző vagy akár a teljes operációs rendszer reszponzivitásának megszűnéséhez vezethet a túlzott memóriaigény miatt. Ezért olyan megoldást kellett találni, amely sem a szervert, sem a kliensgépet nem terheli túl. Ezt úgy tudtuk megoldani, hogy a kétdimenziós térképeket offline legeneráltuk, de a modell méretét úgy korlátoztuk, hogy az ábra megjelenítése és böngészése legalábbis egy nem túl korlátozott memóriakapacitású klienskonfiguráció esetén ne jelentsen gondot.[10] A megjelenítendő modellt a korpuszban legalább háromszor előforduló szófajkóddal annotált lemmák képére korlátozva elfogadható modellméretet kaptunk (13500 lexikai elem). A szemantikus térképek megjelenítését végző kódot, amely a javascript-alapú cytoscape.js gráfvizualizációs és -szerkesztő csomagon alapul (Franz és mtsai, 2015), a Novák és Novák (2020)-ban bemutatott kód adaptálásával készítettük el.

[8] Az esetleges elemzési hibáktól eltekintve ennél a korpuszméretnél a különböző szófajú lemmák szétválasztása egyértelműen jelentős mértékben javítja a modell minőségét, és nem vezet adatritkasági problémákhoz.

[9] Az eredeti modellbeli távolságokkal arányos feltételes valószínűségeket adó gaussi eloszlások és a párdimenziós kép pontjai közötti távolságokat adó Student t-(Cauchy)-eloszlások közötti KL-távolságot optimalizálja. Erre utal módszer nevében a *t*-.

[10] 4GB RAM-mal szerelt laptopon Chrome böngészőben problémamentesen működik.

4. A modellek előállítása

4.1. Az elemzett TMK-n betanított modell

A TMK korpusz a korábbi kísérleteinkhez használt korpusznál három nagyságrenddel kisebb méretű, ezért a korábban alkalmazott módszerek még a lemmatizálással együtt sem adtak a nagy korpuszon kapott modellhez hasonló minőségű eredményt. Valamelyest enyhített a problémán, hogy a modellek létrehozására a fastText eszközt használtuk, amely nem tokenek, hanem karakter-n-gramok reprezentációját hozza létre, így a tanítóanyagban nem szereplő szavakhoz is létre tud hozni reprezentációt a szót alkotó n-gramok reprezentációjának átlagolásával. Emellett kevés minta esetén is viszonylag értelmes eredmény jöhet ki, ha a hasonló szavaknak valóban van közük egymáshoz. Ugyanakkor ez a megközelítés kevés minta esetén reprezentációs problémákhoz vezethet a véletlen hasonlóságok esetében. Pusztán a kb. egymillió szavas elemzett TMK korpuszon betanítva a modell nem volt képes arra, hogy a ritka szóalakokhoz a néhány előfordulásuk alapján megfelelő reprezentációt hozzon létre, ezért az ilyen elemekhez a legközelebbi szomszédok lekérdezésekor leginkább a hasonló karaktersorozatokat tartalmazó, de nyelvileg nem feltétlenül releváns találatok jönnek ki. Gyakoribb szavaknál is sokszor inkább a tematikus, mint a nyelvi hasonlóságok dominálnak (4. és 5. ábra).

0	harc[N]	1	30
1	harc[N]	1.00000005607	30
2	arc[N]	0.8127	15
3	sarc[N]	0.7738	6
4	harc+hely[N]	0.7506	3
5	bérc[N]	0.7011	3
6	harcol[V]	0.6879	5
7	hab[N]	0.6382	6

0	sátor[N]	1	36
1	sátor[N]	1.0000	36
2	Bátor[N]	0.7109	16
3	szenátor[N]	0.6908	4
4	tor[N]	0.6446	9
5	tutor[N]	0.6301	5
6	sátán[N]	0.6213	5
7	sás[N]	0.6201	3

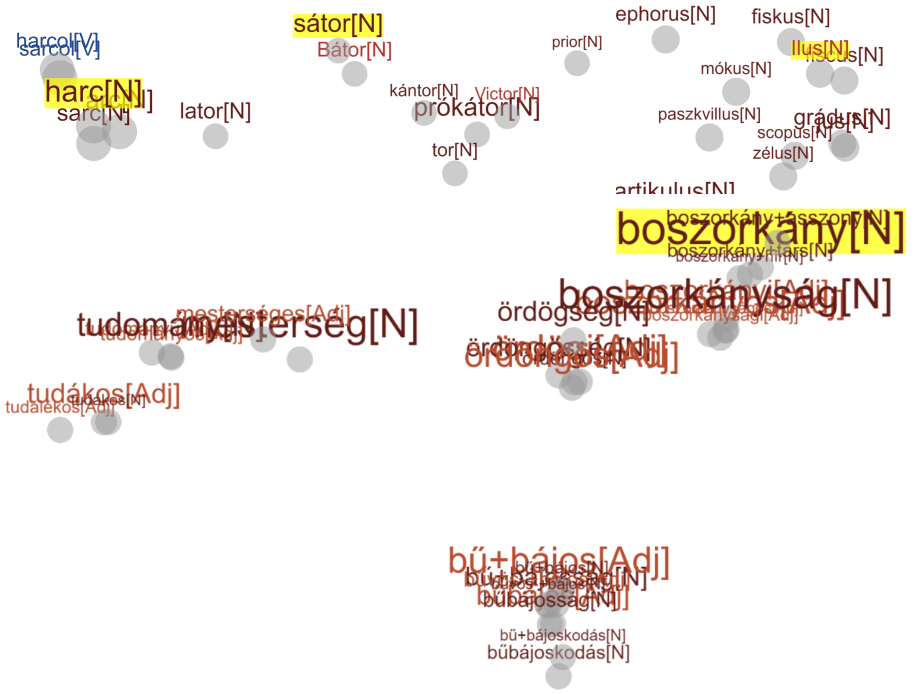
0	öreg[Adj]	1	282
1	öreg[Adj]	1.0000	282
2	öreg[N]	0.8150	10
3	öregség[N]	0.6287	6
4	öreg+leány[N]	0.5572	9
5	öreg+bíró[N]	0.5520	5
6	agg[Adj]	0.5465	3
7	öreg+ember[N]	0.5417	7

0	apa[N]	1	117
1	apa[N]	1.00000002151	117
2	papa[N]	0.7644	5
3	apát[N]	0.7428	3
4	kapa[N]	0.7406	16
5	a[N]	0.7375	9
6	nagy+apa[N]	0.7357	3
7	kupa[N]	0.7251	3

4. ábra. Néhány legközelebbi szomszéd a pusztán a TMK-ból generált modellben.

4.2. Módosított algoritmus

A problémákat úgy próbáltuk orvosolni, hogy további tanítóanyaggal egészítettük ki a korpuszt. Itt azonban problémát jelentett, hogy a hozzáadott tanítóanyagot is a korpusz elemzésével kompatibilis elemzéssel kellett ellátni ahhoz, hogy annotált anyagon alapuló modellt tudjunk létrehozni. Felmerült az az ötlet, hogy az algoritmus módosításával esetleg elemzetlen szöveget is lehetne használni. Ehhez a kísérlethez a fastText CBOW algoritmusának módosított változatát használtuk (CBOW/A), amely alkalmas olyan vektortérmodell létrehozására, amely egyszerre tartalmazza a felszíni szóalakok és az elemzett lemmák reprezentációját (Novák és mtsai, 2019). Az algoritmus alkalmazásához olyan korpuszreprezentációra van szükség, amely a felszíni alakok mellett azok valamilyen annotált



5. ábra. Néhány részlet a TMK-ból generált modellben.

változatát is tartalmazza (1c). Az annotációkat konfigurálható prefix jelöli (a példában: .). A pusztán az elemzéseket tartalmazó modell készítéséhez az eredeti CBOW algoritmus használatakor korábban a (1b)-ben látható formátumot használtuk a modell tanításához.

(1) a. Szeretettel való szolgálatomat ajánlom kegyelmednek, édes szívem!

b. szeretet[N] [Ins] való[Adj] szolgálat[N] [PxS1.Acc] ajánl[V] [S1.Def] kegyelme[N|Pro] [PxS2.Dat] , édes[Adj] szív[N] [PxS1] !

c. Szeretettel .szeretet[N] való .való[Adj] szolgálatomat .szolgálat[N] ajánlom .ajánl[V] kegyelmednek .kegyelme[N|Pro] , édes .édes[Adj] szívem .szív[N] !

Tanításkor az algoritmus véletlenszerűen mintavételezi az egyes korpuszpozíciókban a felszíni szólakot és az adott pozícióhoz tartozó annotációt, így a tanítás során a korpuszon többször végigmenve a felszíni szóalakok és az annotációk reprezentációja is létrejön. A mi konkrét esetünkben a felszíni szóalakok a normalizált alakok, az annotációt pedig a szófajcímkével ellátott lemmák alkotják.

A CBOW/A algoritmust pusztán a TMK-n futtatva a lemmák modellbeli képe nem javult észrevehető módon, így önmagában az algoritmus lecserélése nem

javított a modell minőségén. Abban reménykedtünk azonban, hogy a tanítókorpusz bővítésével az n-gramok jobb reprezentációja segítheti a TMK lemmáinak jobb minőségű ábrázolását is.

4.3. A korpusz bővítése

A korpuszt olyan mai magyar szövegekből vett mondatokkal egészítettük ki, amely olyan szavakat tartalmaz, amelyek a TMK korpuszban is előfordultak, de 100-nál kevesebb előfordulásuk volt. Az új tanítóanyag első változata úgy állt össze, hogy a webkorpuszból szűrt anyagot elemzetlenül adtuk hozzá a TMK elemzett annotált anyagához. A webkorpuszból vett kiegészítés mérete 41,8 millió szó volt. A modell felépítése után azt visszaszűrtük csak a TMK szavaira.

Ebben a modellváltozatban a felszíni szóalakok legközelebbi szomszédait megnézve azt láttuk, hogy azok reprezentációja valóban nagyon sokat javult, mert a TMK-ban ritkább szóalakokra sok példa volt a bővített korpuszban. Azonban visszaszűrve a lemmák reprezentációjára semmilyen lényegi javulást nem láttunk ezek minőségében. Ráadásul a szóalakok reprezentációja nagyon eltávolodott a lemmákétól.

Ezért a következő modellváltozat elkészítéséhez a webes korpuszból vett anyagot is leelemeztük a TMK elemzéséhez használt elemzőlánccal. Ezután az így kapott modellt is visszaszűrtük csak a TMK szavaira. Ebben a modellben a lemmák legközelebbi szomszédait megnézve azt találtuk, hogy a lemmák reprezentációja is elfogadható minőségűre javult (6. és 7. ábra).

0	harc[N]	1	2851
1	harc[N]	1.00000006884	2851
2	csata[N]	0.7949	1661
3	háború[N]	0.7603	2624
4	harc+hely[N]	0.7445	10
5	ellenség[N]	0.7207	2500
6	támadás[N]	0.7203	1754

0	sátor[N]	1	895
1	sátor[N]	1.00000011473	895
2	szoba[N]	0.6728	4290
3	ház[N]	0.6591	16842
4	aszta[N]	0.6455	4927
5	láda[N]	0.6424	908
6	kocsi[N]	0.6417	4385

0	öreg[Adj]	1	3024
1	öreg[Adj]	1.00000006603	3024
2	öreg[N]	0.8053	579
3	vén[Adj]	0.7576	693
4	öreg+ember[N]	0.7264	238
5	idős[Adj]	0.7261	2911
6	szegény[Adj]	0.6960	4075

0	apa[N]	1	5353
1	apa[N]	1.0000	5353
2	anya[N]	0.8621	8722
3	édes+apa[N]	0.8062	1159
4	édes+anya[N]	0.7924	2425
5	nagy+apa[N]	0.7924	644
6	nagy+anya[N]	0.7919	407

6. ábra. Néhány legközelebbi szomszéd a kiegészített korpuszból generált modellben.

Végeztünk egy harmadik kísérletet is, amelyben azt vizsgáltuk meg, hogy van-e jelentősége, hogy menet közben a modellt a TMK-ban nem szereplő lemmákra is betanítjuk. Ebben a kísérletben a tanítóanyag úgy állt elő, hogy a kiegészítő korpuszban a TMK-ban nem szereplő szavak elemzését eldobtuk és így tanítottuk be a neurális modellt. Majd a modellt itt is visszaszűrtük csak a TMK szavaira. A modellbe betekintve azt találtuk, hogy ez a megközelítés az előző változathoz hasonló modellt eredményezett.

4.4. A lexikai térképek

Az így előállt modellekből a kétdimenziós térképeket előállítva azt találtuk, hogy tulajdonképpen a pusztán a TMK-ból a korpusz (1b)-ben látható elemzett alak-

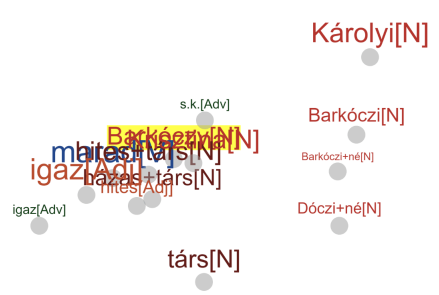
7. ábra. Néhány részlet a bővített korpuszból generált modellben.

jából az eredeti CBOW algoritmussal készített lemma modell is jól használható áttekintést ad a korpusz szavairól, amelyben azonban nem elsősorban a nyelvi jellegű szerveződési szintek alapján csoportosulnak az elemek, hanem sok helyen inkább a korpuszra jellemző tematikus csoportok dominálnak. A korpusz anyagát ismerő kutatók számára hamar feltűnnek azok a sűrűsödési pontok a térképen, ahol az egyes részkorpuszok nagyon jellemző fordulatai, nevei csoportosulnak (8. ábra). Emellett helyenként pusztán a sztringhasonlóság hozza egymáshoz közel az elemeket, amely ugyan az esetek nagy részében az egymáshoz közel lévő elemek nyelvi hasonlóságával jár együtt, de időnként egyszerűen csak rímelnek az egymáshoz közel lévő szavak hasonlóan az 5. ábrán felül látható esetekhez(pl. *arc-harc-sarc*).

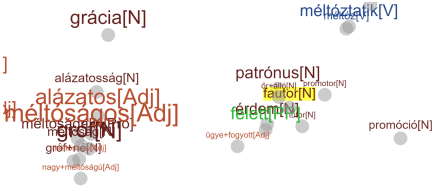
A kiegészített modellekből készült térképeken sokkal inkább a lexikai tér nyelvi szerveződése érhető tetten. A különböző szófajú elemek nem keverednek olyan mértékben egymással, mint a kis modellből készített térképen. Helyenként tetten érhető, hogy a mai korpuszban domináló jelentések irányába mozdult el a kép (pl. míg a pusztán a TMK-ból készített térképen a *mesterséges* és a *tudományos* boszorkánysághoz köthető fogalmakként jelennek meg, a *közösködik* a *közösül* szinonimája, a kiegészített korpuszból készült modellből generált térképen ezeknek a lexémáknak a képe elmozdult a mai jelentésüknek megfelelő helyekre.)

A térképeken időnként meglepő helyeken jelennek meg meglepő lexikai elemek. Ezeknek a jelenségeknek könnyen utána járhatunk az adott elemre kattintva kapott lekérdezések eredményére rátekintve, és azt találjuk, hogy az elemzett korpusz kézzel nem ellenőrzött részéből származó reprezentációkról van szó (9. ábra). Például a női nevek között feltűnő *bután* főnév az egyik boszorkányper-

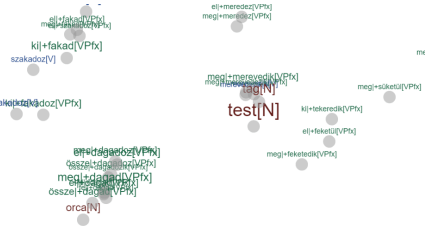
(a) A jogászszereplőket leíró jellegzetes fordulatok szavai a perszövegek formulaszerű hivatalos részeiből.



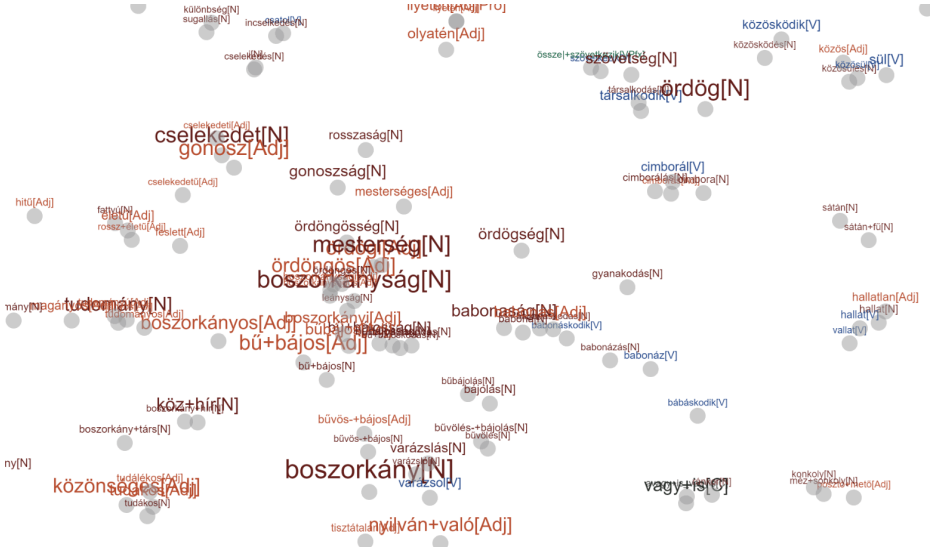
(b) A Károlyi-Barkóczi-levelezés jellegzetes fordulatainak elemei.



(c) A peregrinuslevelek jellegzetes szavai.



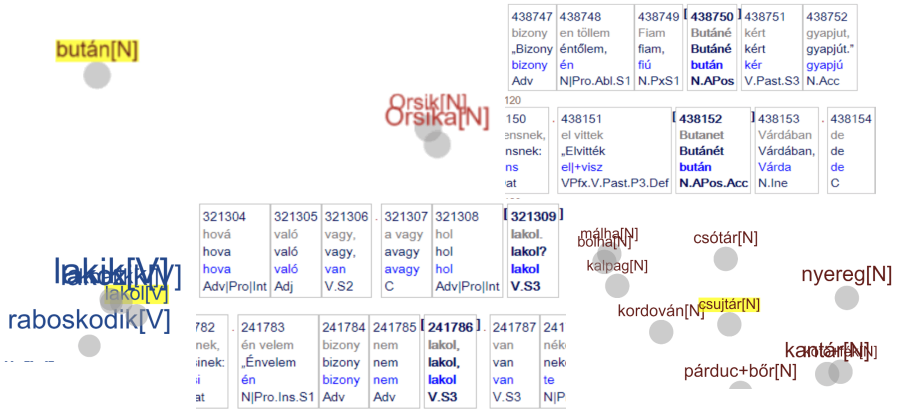
(d) Rontások hatásai a boszorkányperekből, és amit érint.



(e) A boszorkányperek jellegzetes szavai.

8. ábra. Néhány tematikus sűrűsödési pont a TMK-ból az eredeti CBOW algoritmussal generált modellben.

szereplő, Butáné ‘vicces’ elemzéséből adódik, hasonlóan a *lakol* közelsége a *lakik*-hoz jól mutatja, hogy valójában nem a *(meg)lakol* igéről van szó, hanem a lakik szubsztenderd második személyű alakjáról, amely néhány ellenőrizetlen szövegben hibás elemzéssel maradt benne. A modell azokat az eseteket is felszínre hozza, ahol a normalizálás során a fonológiai variabilitásból adódó különbségeket nem sikerült teljesen semlegesíteni (pl. *csujtár-csótár* nyereg alatti lótakaró’).



9. ábra. Néhány példa elemzési/normalizálási hibákra. Az elem helye a térképen utal arra, hogy hogyan kell javítani.

5. Összefoglalás

Cikkünkben a Történeti Magánéleti Korpusz (TMK) webes lekérdezőfelületén elérhető interaktív tematikus-szemantikus lexikai térképet mutattuk be a kereső egyéb újdonságai mellett. A pusztán a TMK-ból készített, a korpusz kis mérete miatt jellegében inkább tematikusnak, mint igazán nyelvinek mondható szóbeágyazási modell mellett a TMK kibővítésével nyert korpuszból készített már inkább nyelvi-szemantikus modellekből a t-SNE algoritmussal nyert kétdimenziós lexikai térképek elemeire kattintva közvetlenül is indítható az adott nyelvi elemre vonatkozó korpuszlekérdezés. A térképek ugyanakkor a szövegek gépi feldolgozásakor, illetve kézi ellenőrzésekor bent maradt hibákra is felhívják a figyelmet, könnyítve ezzel a hibajavítást.

Köszönetnyilvánítás

Jelen kutatás a Nemzeti Kutatási Fejlesztési és Innovációs Alap által biztosított támogatással a K 116217 számú projekt, illetve a K 15 pályázati program keretében valósult meg.

Hivatkozások

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606 (2016)

Dömötör, A., Gugán, K., Novák, A., Varga, M.: Kiútkeresés a morfológiai labirintusból : korpuszépítés ó- és középmagyar kori magánéleti szövegekből. Nyelvtudományi Közlemények 113, 87–114 (2017)

Franz, M., Lopes, C.T., Huck, G., Dong, Y., Sümer, S.O., Bader, G.D.: Cytoscape.js: a graph theory library for visualisation and analysis. In: Bioinformatics (2015)

van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)

Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. CoRR abs/1301.3781 (2013), `http://arxiv.org/abs/1301.3781`

Novák, A.: Milyen a jó Humor? In: I. Magyar Számítógépes Nyelvészeti Konferencia. pp. 138–144. SZTE, Szeged (2003)

Novák, A., Laki, L.J., Novák, B.: CBOW/A: módosított CBOW algoritmus annotált szövegekből készített vektortérmodellek létrehozására. In: XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 37–48 (2019)

Novák, A., Novák, B.: Magyar szóbeágyazási modellek kézi kiértékelése. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia : MSZNY 2018. pp. 67–77 (2018)

Novák, A., Novák, B.: Bu-Bor-éK: grafikus címkenormalizáló eszköz. In: XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 303–312 (2020)

Novák, A., Novák, B., Wenszky, N.: Szóbeágyazási modellek vizualizációjára és böngészésére szolgáló webes felület. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017). pp. 355–362 (2017)

Novák, A., Wenszky, N.: Ó- és középmagyar szóalaktani elemző. In: IX. Magyar Számítógépes Nyelvészeti Konferencia [Ninth Hungarian Conference on Computational Linguistics]. pp. 170–181 (2013)

Orosz, Gy., Novák, A.: PurePos 2.0: a hybrid tool for morphological disambiguation. In: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013). pp. 539–545. Incoma Ltd. Shoumen, Bulgaria, Hissar, Bulgaria (2013)

Petersen, U.: Emdros — a text database engine for analyzed or annotated text. In: In: Proceedings of COLING 2004. (2004) 1190–1193 (2004)